Capstone Introduction

Process, Timeline & Tips

Amber Yandow
052620-PT-DS

// FLATIRON SCHOOL

Agenda

1. Data Science Skills/Jobs
   a. Where our grads get jobs
   b. Hard/Soft skills of those jobs

2. Agile methodology
   a. Daily Stand-ups
   b. Trello boards
   c. Weekly 1:1s

3. Project Requirements and examples

Which "Data Profession" titles exist?

DS vs DA

SPECIALIZED SKILLS	DATA SCIENTIST n = 19577	DATA ANALYST / 49234
Python	76.9%	21.3%
Machine Learning	67.0%	5.8%
SQL	51.0%	52.9%
Predictive Models	26.3%	3.0%
Big Data	24.8%	6.6%
Apache Hadoop	23.0%	4.7%
Data Mining	22.8%	9.8%
R	21.9%	5.9%
Artificial Intelligence	21.7%	1.5%
Java	20.5%	4.2%
Tableau	19.6%	25.2%
SAS	19.5%	12.6%
Natural Language Processing	19.3%	1.0%
Deep Learning	18.4%	1.0%
Data Visualization	17.7%	11.9%
Economics	16.7%	8.4%
Data Management	6.6%	17.1%
Data Quality	4.7%	16.8%

DS vs DA

BASELINE SKILLS	DATA SCIENTIST n = 19577	DATA ANALYST / n = 49234
Microsoft Excel	0.085	0.418
Research	0.429	0.231
Communication Skills	0.348	0.404
Problem Solving	0.23	0.257
Detail-Oriented	0.078	0.233
Teamwork / Collaboration	0.301	0.231
Writing	0.136	0.186
Microsoft Office	0.033	0.16
Organizational Skills	0.059	0.144
Creativity	0.146	0.093
Analytical Skills	0.07	0.126
Microsoft Powerpoint	0.031	0.125
Planning	0.066	0.112
Written Communication	0.084	0.105
Presentation Skills	0.114	0.098

Remember Kaggle's state of ML & DS in 2020

Congratulations!

Capstone Project Technical Requirements, student examples and advice

Stand-up

Tuesdays SGs will be
stand-ups to discuss
your progress

Instructors acting as
"managers" overseeing
the process and taking
notes

Flatiron DS Stand-up "Script"

Project Elevator Pitch (30 seconds, should explain what problem/need you are addressing, and how you are addressing it)
_____
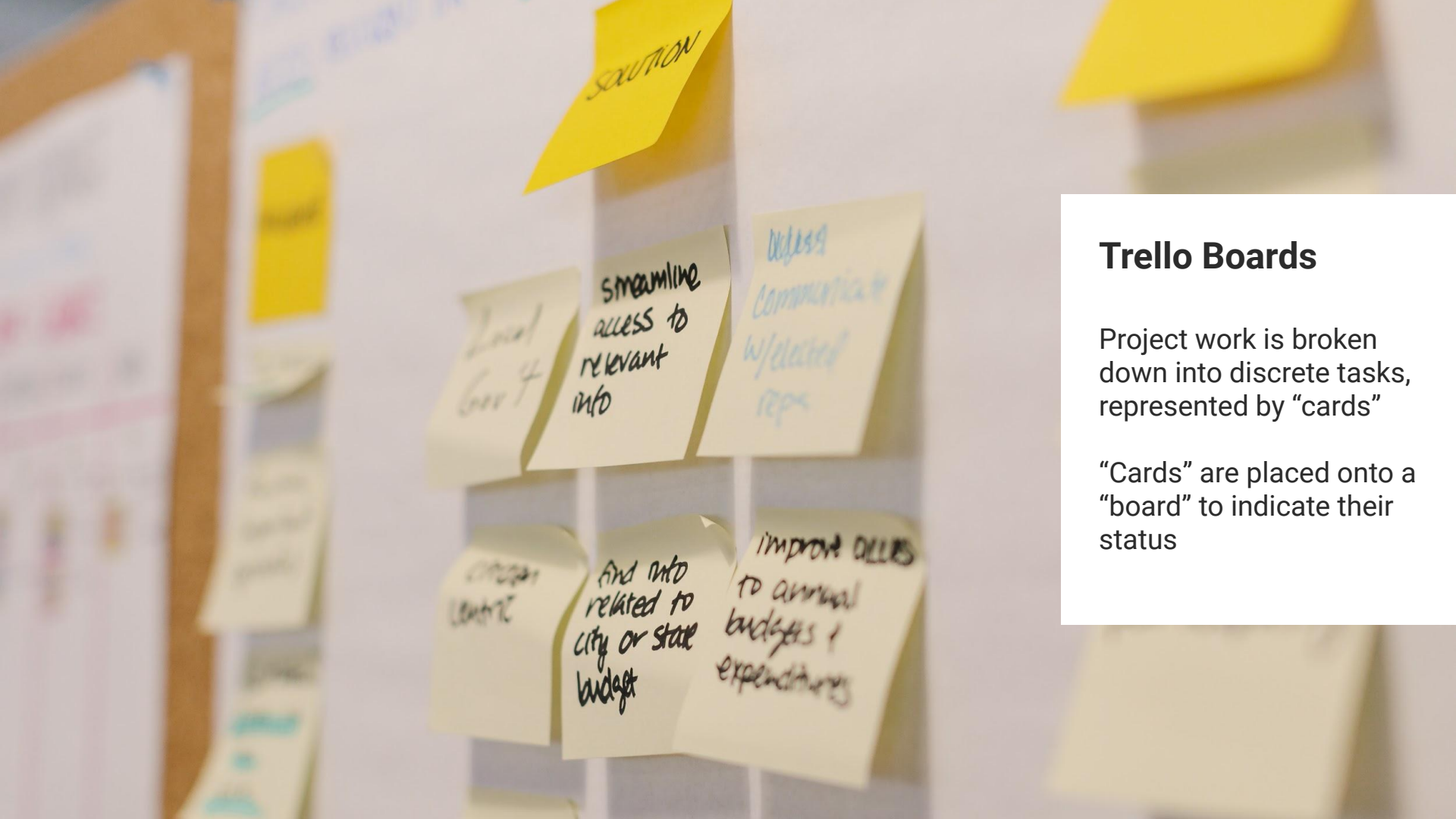_____

Accomplishments
Since our last check-in, I have _____

Plans for Today
Today I plan to _____

Blocking issues (if relevant)
I need help with _____

Trello Boards

Project work is broken down into discrete tasks, represented by "cards"

"Cards" are placed onto a "board" to indicate their status

Flatiron DS Trello Board Format

ICE BOX	TO DO	DOING	DONE

Nice-to-have tasks that you might have time for during capstone, might work on after graduation, or might never get to, e.g.:
- Adding interactive visualizations
- Deploying live website
- Conducting a user study with surveys

Everything for your MVP starts here, unless you have already completed it before capstone started, e.g.
- Data collection
- Writing repository README(s)
- Creating presentation slides

A card should only stay in this category for a day or two at most

Instructors should be able to look here to get an immediate snapshot of what you're working on and what you might need help with

Pretty self-explanatory… put cards here when the task has been completed!

Trello Board Tools

- [Trello](#) tends to be most popular
- Connect your Trello account to Slack [HERE](#)
- We will use this as a guide during our weekly 1:1s

Element	Complete Project	Minimum Viable Product (MVP)	Incomplete Project
Business Understanding	Clearly explains the **real-world value** the project has **for a specific stakeholder**.	**Clearly explains the real-world problem** that the project sets out to solve.	Project has **unclear goals** (e.g. explore a dataset) or **no real-world relevance**.
Data Understanding	**Relates data source and properties of variables to the real-world problem** of interest.	**Describes data source and properties of all variables** used in data preparation and modeling.	**Does not describe data source or explore variables** used in data preparation or modeling.
Data Preparation	Data preparation is fully **documented, including valid justification** for decisions	All data preparation steps are **reproducible and justifiable**.	Data preparation is **not reproducible or justifiable** given the problem of interest.
Modeling	Model development is correct, iterative, and fully **documented, including valid justification** for decisions	Models are **developed iteratively and justifiably**, proceeding from a simple baseline model to more complex models.	Models are **not developed iteratively or justifiably** given the problem of interest.
Evaluation	Clearly **explains how well the project solves the real-world problem** of interest.	Cross validation is used correctly to **evaluate model performance**.	Cross validation is **not used or is used incorrectly** to evaluate model performance.
Readme Content	README is **error-free** and **well-written**: clear, concise, complete, organized, narrative, starting with a project overview.	README **correctly includes all required elements**: data science process steps, future improvement ideas, repository navigation, reproduction instructions, links to presentation and sources.	README **omits required elements or has substantial errors** in writing or substance.
Notebook Content	Notebook is **error-free** and **well-written**: clear, concise, complete, organized, narrative, starting with a project overview.	Notebook **correctly includes all required elements**: data understanding, data preparation, modeling, and evaluation.	Notebook **omits required elements or has substantial errors** in writing or substance.
Presentation Content	Presentation is **error-free** and **well-written**: clear, concise, complete, organized, narrative, starting with a project overview.	Presentation **correctly includes all required elements**: introduction, data science process steps, future improvement ideas, and contact info.	Presentation **omits required elements, or has substantial errors** in writing or substance.
Presentation Style	**All** presentation slides have a professional style.	**Most** presentation slides have a professional style: uncluttered, light on text, no unnecessary jargon, visuals clearly demonstrate key points.	Most slides have **unprofessional style**: cluttered, text- or jargon-heavy, visuals that are dense, unclear or unnecessary
Sourcing	Properly cites all content created by others and **provides as much access as feasible and permissible**.	Properly **cites all content created by others** (e.g. data, code, images).	Has **uncited content created by others**, or **uses content without appropriate permission**.